WEAKLY SUPERVISED HMM LEARNING FOR SPOKEN WORD ACQUISITION IN HUMAN COMPUTER INTERACTION WITH LITTLE MANUAL EFFORT

Meng Sun[1], *Hugo Van hamme*[2], *Xiongwei Zhang*[1]

[1]Lab of Intelligent Information Processing, PLA University of Science and Technology, Nanjing, China 210007
[2]Department of Electrical Engineering-ESAT, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, Leuven, Belgium B-3001

sunmengccjs@gmail.com, hugo.vanhamme@esat.kuleuven.be, xwzhang9898@163.com

ABSTRACT

In this paper, weakly supervised HMM learning is applied to modeling word acquisition towards human-computer interaction with little manual effort. The only imposed supervisory information is initializing the learning algorithms by two labeled data samples per pattern. Experiments on TIDIGITS show that our recently proposed algorithm, Baum-Welch learning regularized by non-negative Tucker decomposition, succeeds in finding good solutions in the sense of yielding high recognition accuracy on the testing data which approximate the supervised baseline (98.0% vs 98.9%).

Index Terms— spoken word learning, hidden Markov models, semi-supervised learning, non-negative matrix factorization, regularization

1. INTRODUCTION

Speech is becoming an important way of communications between human and computers, e.g. the Siri system in iPhone. However, the development of such a speech recognition device requires sophisticated prior knowledge and careful coding of the language. Once new words are emerging or the language is not resourceful, the system has to be coded again. That is the learning process of computers is not as efficient as human especially when tackling unseen patterns. To make computer to be able to learn words in an autonomous way, computational modeling of spoken word learning or vocabulary acquisition has attracted a lot of attention in the past few years [1]∼[4]. The critical problem to be solved in the task is how to acquire *accurate* speech representations by using *little* supervision. Being different from the fully supervised learning in automatic speech recognition (ASR) where word-by-word transcriptions are available, computational modeling of word acquisition is actually a *semi-supervised learning* problem. It is also different from the out-of-vocabulary (OOV) detection problem since in OOV only a few "slots" are not modeled by the learning agent, while in word acquisition most labels of the data are missing. Therefore, computational spoken word acquisition is a semi-supervised learning with only *weak labels*. In an extreme case without any labeled data, spoken word learning can also be related to unsupervised spoken pattern discovery [5][6]. Various speech representation methods have been explored towards modeling of spoken word acquisition, such as hidden Markov models (HMM) in [3], concept matrices in [4], clusters of segment traces in [5][6][7] and non-negative matrix factorization (NMF) in [2][8].

HMMs are the conventional tools for modeling speech in state-of-the-art ASR systems. However, to yield good recognition performance, those HMMs should be trained by a sufficiently large amount of *labeled* data. HMMs with unsupervised or semi-supervised learning are easily trapped at poor local optima which may not show strong relations to the underlying words. Besides the Baum-Welch (BW) algorithm [9], other ways are explored for training HMMs, such as the algorithm of learning HMM parameters from bags-of-co-occurrences of the observation sequences in [10][11][12]. Experiments on artificial data show that the algorithms can indeed learn the HMM parameters.

Aiming at unsupervised spoken pattern discovery, we proposed a new unsupervised training paradigm for HMM learning by combining non-negative Tucker decomposition (NTD) and BW in [13]. In the current paper, we will further improve the NTD regularized BW algorithm of [13], and apply it on semi-supervised word learning with weak labels where the only supervisory information is using a few labeled samples per word as initialization of the HMM. The semi-supervised baseline trained by BW and the supervised baseline will also be provided for comparison.

The organizations of the paper are as follows. The configuration of the HMM is presented in Section 2. The algorithm of NTD regularized BW is briefly introduced in Section 3. The experiments and results are reported in Section 4. In Section 5, we give out the conclusions based on the results.

978-1-4799-2186-7/14/$31.00 ©2014 IEEE

2. HMMS FOR WORD LEARNING

2.1. Problem definition and HMM configuration

In the spoken word learning problem, we are given a set of utterances in the form of observation sequences,

$$\mathbf{O}^{(n)} = \{O_1^{(n)}, \ldots, O_{T_n}^{(n)}\} \tag{1}$$

where $\mathbf{O}^{(n)}$ denotes the sequence containing one or more words, n is the index of the utterance, $O_t^{(n)}$ is the observation symbol at frame t which can be obtained by vector quantization using a pre-trained codebook [8], T_n is the number of frames in utterance n. The codebook is a collection of codewords in the space of the observation symbols: $\mathbf{v} = \{v_1, \ldots, v_M\}$.

An HMM with the following configuration is designed for modeling the spoken words and their transition probabilities.

- A set of sub-HMMs $\{\mathbf{\Lambda}^{(r)}\}_{r=1}^{R}$ each of which models a word. The parameters involved in the sub-HMM $\mathbf{\Lambda}^{(r)}$ are its emission matrix $\mathbf{B}^{(r)}$ and its transition matrix $\mathbf{A}^{(r)}$. R is the total number of words.
- The initial word distribution $\boldsymbol{\pi}$ which has size $R \times 1$. π_r denotes the probability that word r is the start of an utterance.
- The transition matrix between the words or sub-HMMs, $\mathbf{T}$ with size $R \times R$, where $T_{r,r'}$ is the conditional probability $\Pr(\mathbf{\Lambda}^{(r')}|\mathbf{\Lambda}^{(r)})$. The matrix actually models a bigram grammar of the words.

Let $\mathbf{\Lambda}$ denote the HMM with the above parameters:

$$\mathbf{\Lambda} := \{\{\mathbf{A}^{(r)}, \mathbf{B}^{(r)}\}_{r=1}^{R}, \mathbf{T}, \boldsymbol{\pi}\}. \tag{2}$$

The word learning problem thus boils down to the estimation of the parameters in the HMM $\mathbf{\Lambda}$.

2.2. Semi-supervised learning with weak labels

Expectation maximization (EM) methods, such as segmental k-means [14] and Baum-Welch (BW) which maximize the likelihood of the data in Eq.(3), are usually applied to estimate the unknown parameters of an HMM.

$$\sum_{n=1}^{N} \log\Pr(\mathbf{O}^{(n)}, \mathbf{L}^{(n)}|\mathbf{\Lambda}) \tag{3}$$

Often, the sequential data $\mathbf{O}^{(n)}$ comes with sequential labels $\mathbf{L}^{(n)}$, resulting in a *supervised* training problem. For instance, in ASR the training data is labeled in terms of the words from which an HMM representation for each word label is learned from the data. However, in the task of computational vocabulary acquisition, no prior language knowledge is assumed to be known by the learning agent. Therefore, the labels of the sequences are not provided, or only a few labels are provided. Without loss of generality, we assume the first N_1 sequences are labeled where $0 \leq N_1 \leq N$. The HMM training is therefore *semi-supervised* in Eq.(4).

$$\sum_{n=1}^{N_1} \log\Pr(\mathbf{O}^{(n)}, \mathbf{L}^{(n)}|\mathbf{\Lambda}) + \sum_{n=N_1+1}^{N} \log\Pr(\mathbf{O}^{(n)}|\mathbf{\Lambda}) \tag{4}$$

With the assumption of weak labels, we consider the case that only a few utterances are labeled, i.e. $N_1 \ll N$, so the supervised term in Eq.(4) contributes little to the total objective function. Certainly one can adjust the weights of the two terms in Eq.(4), e.g. using the approach proposed in [15]. However, in this paper, we investigate a case with even weaker supervision, where the labeled data is only utilized for initialization.

3. NTD REGULARIZED BW

In this section, we briefly introduce the algorithm of NTD regularized BW.

3.1. Learning HMM from co-occurrence statistics

The sub-HMM $\mathbf{\Lambda}^{(r)}$ can be identified from co-occurrence statistics of its observations. Let $\mathbf{C}^{(r)}$ is a $M \times M$ matrix storing the co-occurrences of observation symbols, i.e. the element $C_{m,m'}^{(r)}$ is the probability of observing $(v_m, v_{m'})$ in the sequences generated by the sub-HMM $\mathbf{\Lambda}^{(r)}$. Therefore, if $\mathbf{C}^{(r)}$ is available, $\mathbf{A}^{(r)}$ and $\mathbf{B}^{(r)}$ can be learned by non-negative matrix tri-factorization (NMTF):

$$\mathbf{C}^{(r)} := \mathbf{B}^{(r)}\mathbf{A}^{(r)}(\mathbf{B}^{(r)})^T, \tag{5}$$

Actually $\mathbf{C}^{(r)}$ can be estimated as follows. An utterance is first represented by its co-occurrences of observation symbols by a transform called *histogram of acoustic co-occurrences* (HAC) in [8]: $\mathbf{X}^{(n)} := \text{HAC}(\mathbf{O}^{(n)})$, where $X_{m,m'}^{(n)} = \sum_{t=1}^{T_n} \delta(O_t^{(n)} = v_m, O_{t+1}^{(n)} = v_{m+1})$. If the utterance contains several words (i.e. generated by their sub-HMMs), $\mathbf{X}^{(n)}$ is a mixture of the co-occurrences of observations of the corresponding sub-HMMs. That is,

$$\mathbf{X}^{(n)} \approx \sum_r \mathbf{C}^{(r)} H_{r,n}, \tag{6}$$

where $H_{r,n}$ is the frequency of appearance of the word or sub-HMM r in utterance n, or the weight of word r in mixture (6). $\mathbf{H}$ is thus called the weight matrix of the sub-HMMs.

In the NTD learning of HMMs, we first learn the $\mathbf{C}^{(r)}$'s from the input sequences $\mathbf{O}^{(n)}$. Non-negative matrix factorization (NMF) can serve this goal. Non-negative matrix tri-factorization (NMTF) is subsequently applied on Eq.(5) to learn each of the sub-HMMs. The initial word distribution $\boldsymbol{\pi}$ and the grammar $\mathbf{T}$ are not involved in NTD and they should be updated by using conventional algorithms like BW.

3.2. BW training with NTD regularization

The new objective function in NTD regularized BW is,

$$\max_{\mathbf{\Lambda},\mathbf{H}} \quad \sum_n \log\Pr(\mathbf{O}^{(n)};\mathbf{\Lambda}) \\ -\lambda * \mathrm{KLD}(\mathbf{X}^{(n)} || \sum_r \mathbf{B}^{(r)}\mathbf{A}^{(r)}(\mathbf{B}^{(r)})^T H_{r,n}), \tag{7}$$

where the first term is the log-likelihood of the observation sequence $\mathbf{O}^{(n)}$ on the HMM $\mathbf{\Lambda}$ and the second term is the Kullback-Leibler divergence (KLD) of the co-occurrence matrix $\mathbf{X}^{(n)}$ and its reconstruction from the sub-HMMs. The prospective HMM parameter estimates should fit for both data representations: the first term measures how well the HMM is able to generate the actual sequences, while the regularization term expresses that the global co-occurrence statistics can be decomposed into additive parts where each part corresponds to a sub-HMM.

$\mathbf{H}$ only occurs in the regularization term, so its update algorithm remains unchanged from the NMF algorithm to solve Eq.(6). $\boldsymbol{\pi}$ and $\mathbf{T}$ only occur in the HMM term so the conventional BW algorithm can update them. The updating of $\mathbf{A}^{(r)}$ and $\mathbf{B}^{(r)}$ involves both BW and NTD and turns out to be the following weighted sums:

$$B^{(r)}_{m,k} = \frac{\mathrm{E}^{(\mathrm{BW})}(n_{m,k,r}) + \lambda \mathrm{E}^{(\mathrm{NTD})}(n_{m,k,r})}{\sum_{m'} \mathrm{E}^{(\mathrm{BW})}(n_{m',k,r}) + \lambda \mathrm{E}^{(\mathrm{NTD})}(n_{m',k,r})} \tag{8}$$

$$A^{(r)}_{k,l} = \frac{\mathrm{E}^{(\mathrm{BW})}(n_{k,l,r}) + \lambda \mathrm{E}^{(\mathrm{NTD})}(n_{k,l,r})}{\sum_{l'} \mathrm{E}^{(\mathrm{BW})}(n_{k,l',r}) + \lambda \mathrm{E}^{(\mathrm{NTD})}(n_{k,l',r})} \tag{9}$$

where $\mathrm{E}^{(\mathrm{BW})}(n_{m,k,r})$ and $\mathrm{E}^{(\mathrm{NTD})}(n_{m,k,r})$ denote the expected number of observations of symbol v_m in state k of sub-HMM $\mathbf{\Lambda}^{(r)}$ and $\mathrm{E}^{(\mathrm{BW})}(n_{k,l,r})$ and $\mathrm{E}^{(\mathrm{NTD})}(n_{k,l,r})$ denote the expected number of transitions from state k to state l of sub-HMM $\mathbf{\Lambda}^{(r)}$ from BW and NTD correspondingly.

The terms $\mathrm{E}(n_{m,k,r})$ and $\mathrm{E}(n_{k,l,r})$ in NTD are the unnormalized estimates of $B^{(r)}_{m,k}$ and $A^{(r)}_{k,l}$ respectively. In this paper, we modify the scaling schemes of the algorithm of NTD regularized BW presented in [13] to keep a balance on the scales of the estimates from NTD and BW. As will be shown below, the scales of the NTD estimates are equal to the number of frames which is equal to the scale of the BW updates [16]. In Eq.(6), the updating algorithm of $\mathbf{C}^{(r)}$ is computed by,

$$C^{(r)}_{m,m'} \leftarrow C^{(r)}_{m,m'} \sum_n \frac{X^{(n)}_{m,m'}}{\sum_t C^{(t)}_{m,m'} H_{t,n}} H_{r,n}. \tag{10}$$

It is straightforward to check that

$$\begin{aligned} \sum_r C^{(r)}_{m,m'} &= \sum_r C^{(r)}_{m,m'} \sum_n \frac{X^{(n)}_{m,m'}}{\sum_t C^{(t)}_{m,m'} H_{t,n}} H_{r,n} \\ &= \sum_n X^{(n)}_{m,m'}, \end{aligned} \tag{11}$$

which states that the count of the co-occurrence $(v_m, v_{m'})$ from all the utterances, $\sum_n X^{(n)}_{m,m'}$, is reallocated to the R sub-HMMs while the total count is retained.

In the NMTF algorithm to estimate $\mathbf{A}^{(r)}$ and $\mathbf{B}^{(r)}$ by factorizing $\mathbf{C}^{(r)}$ in Eq.(5), $\mathbf{A}^{(r)}$ is updated by,

$$A^{(r)}_{k,l} \leftarrow A^{(r)}_{k,l} \sum_{m,m'} \frac{C^{(r)}_{m,m'}}{(B^{(r)}A^{(r)}(B^{(r)})^T)_{m,m'}} B^{(r)}_{m,k} B^{(r)}_{m',l}. \tag{12}$$

It is straightforward to check that $\sum_{k,l} A^{(r)}_{k,l} = \sum_{m,m'} C^{(r)}_{m,m'}$. Therefore the total count of state transitions $\sum_r \sum_{k,l} A^{(r)}_{k,l}$ is equal to $\sum_r \sum_{m,m'} C^{(r)}_{m,m'}$ which is further equal to the total count of observation co-occurrences $\sum_n \sum_{m,m'} X^{(n)}_{m,m'}$ (by Eq.(11)) which is equal to the number of frames in the training data.

The update of $\mathbf{B}^{(r)}$ is given by,

$$\begin{aligned} B^{(r)}_{m,k} \quad \leftarrow \frac{B^{(r)}_{m,k}}{2} \sum_{m',k'} \Big(& \frac{C^{(r)}_{m',m}}{(B^{(r)}A^{(r)}(B^{(r)})^T)_{m',m}} B^{(r)}_{m',k'} A^{(r)}_{k',k} + \\ & \frac{C^{(r)}_{m,m'}}{(B^{(r)}A^{(r)}(B^{(r)})^T)_{m,m'}} B^{(r)}_{m',k'} A^{(r)}_{k,k'} \Big). \end{aligned} \tag{13}$$

It is again straightforward to show that $\sum_{m,k} B^{(r)}_{m,k} = \sum_{m,m'} C^{(r)}_{m',m}$. Thus the total count of observations of the states in all the sub-HMMs, $\sum_r \sum_{m,k} B^{(r)}_{m,k}$, is equal to number of co-occurrences of observations in $\sum_r \sum_{m,m'} C^{(r)}_{m,m'}$ which is equal to the number of frames in the training data as explained above.

4. EXPERIMENTS AND RESULTS

In this section, we test the performance of the algorithm on word learning from TIDIGITS which has a vocabulary of 11 English digits in 8438 training utterances and 1001 test utterances in continuous speech from multiple male and female speakers. Like in conventional ASR systems, speech is chopped into overlapping frames to compute the short-term spectral features known as Mel-Frequency Cepstral Coefficients (MFCC) plus the frame's log-energy. The window length is 25 ms with a frame shift of 10 ms. For each frame, a 12 dimensional MFCC vector is extracted from a bank of 30 Mel-scaled filters. First and second order differences are computed and concatenated to form a 39-dimensional feature vector. A Gaussian mixture of 1000 components is trained without supervision on the training set using maximum likelihood EM training. The Gaussian identities will later be utilized as discrete observable symbols.

4.1. Semi-supervised learning for computational word acquisition

The task is to learn words with a few labeled samples (i.e. utterances containing a single word) which is operated as follows:

- **Teaching:** The HMM is initialized by two randomly-chosen samples for each word. One from a male speaker and the other is from a female speaker. The utterance is uniformly cut into K pieces to initialize the K states of the sub-HMM of the corresponding digit. The initializations from the male and female speakers of the same digit are accumulated. *This process simulates the process where humans teach the computer the vocabulary.*

- **Self-learning:** With the above initializations, the HMM is trained using BW or NTD regularized BW on unlabeled continuous speech which contains the above words. *This process means that the computer does self-learning or reviewing of the words taught before by the teachers to refine the knowledge.*

- **Examination:** The learned HMMs are evaluated by recognizing the words in unseen data. *This process corresponds to the examination on the words contained in the training data.*

4.2. Results and discussions

For the configuration of the HMM, we set R=12 sub-HMMs each of which has K=10 states. The observation alphabet is the set of M=1000 Gaussians. The number of EM passes is 10 for all the algorithms.

The test set contains 3257 words in 1001 utterances. For each learned HMM, we report the lowest word error rate (WER) obtained by optimizing over the word entrance penalty (WEP) using line search from -500 to 0 with a step-size of 10. The choice of utterances for making the initializations could play a role in the learning problem. We therefore conduct 10 initializations and analyze the results statistically. The mean values using the BW algorithm and NTD regularized BW with different regularization parameters are shown in Figure 1. The supervised baseline is also given for reference. From the figure, it is straightforward to see that NTD regularized BW with $\lambda = 1$ improves BW by approximating the supervised baseline. Large λ has the risk of ruining the HMM learning [1] since the NTD learning only cares about the global co-occurrence statistics, not the real sequences. In Section 3.2, we have adjusted the scales of the BW update and the NTD update to be equal to each other. Therefore $\lambda = 1$ would be a reasonable choice which does not favor one update over the other. However, λ's with smaller scales also show good performance as is also witnessed in Figure 1.

Figure 2 shows the relative improvements and the standard deviations over the 10 initializations. The superiority of NTD regularized BW with $\lambda \in [10^{-50}, 1]$ over BW is observed with statistical significance on this task. The difference between the NTD regularized BW algorithm with $\lambda = 1$ and the supervised baseline is not very significant though the supervised baseline is indeed a bit better.

[1] For λ=10, an accuracy of 95.62±1.83 is found, which fails to improve over BW. Because of the scale of Figure 1, this is not shown.

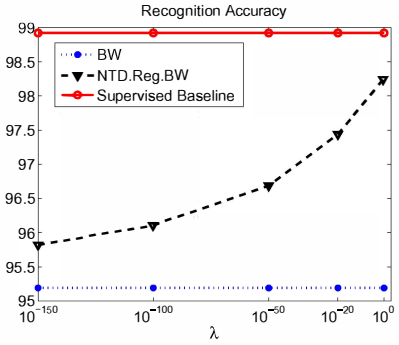

Fig. 1. Comparison of BW and NTD regularized BW on semi-supervised word learning with weak labels. The supervised baseline is also given for comparison. [1]

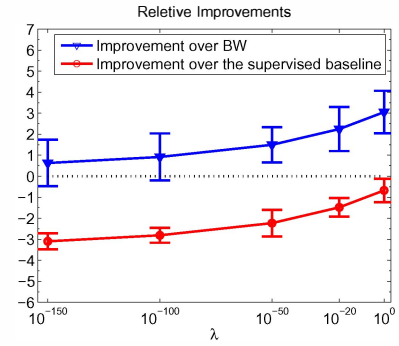

Fig. 2. The relative improvements of NTD regularized BW over the baselines. The performance of NTD regularized BW with $\lambda \in [10^{-50}, 1]$ outperforms BW significantly.

We have shown that a few discrete utterances with labels are important for word learning, compared to the completely unsupervised learning of [13]. Interestingly, a similar phenomenon for human learning was reported in the experiments of [17], where the words' statistical properties were detected successfully by infants only when the infants were exposure to *a combination of continuous speech and discrete utterances.* Notice that as pointed out in [18], there are large similarities between NMF-based co-occurrence learning in [18] and this paper and statistical learning in babies [17]. Brief exposure to *continuous speech only* was not sufficient for accurate learning.

Similar experiments were conducted by [8], [4] and [3] where the full 6214 utterances in the testset of TIDIGITS were used for evaluation. By using *all* the labeled training utterances, the recognition rates of 94.43% and 92.77% were observed in [8] and [4] respectively. By comparing with these results, it is straightforward to see that our algorithm can learn HMM with good performance by only using a few labeled samples. [3] reported 75% recognition rate by only using two

labeled samples, without using any other training utterances. Our results showed that the unlabeled data can be helpful to improve the performance of the model trained on little data.

5. CONCLUSION

The performance of semi-supervised word acquisition with weak labels has been improved by using NTD regularized BW. With only two labeled samples as supervision, the semi-supervised learning yields comparable performance as the supervised baseline. Like the learning process that a student learns words, *teaching*, *self-learning* and *evaluation* are modeled for a computer. However, one stage that is similar to the interactions between students and teachers, is missing which can be called *asking*. This stage is critical for the correction or calibration of the learned knowledge by computers, which would be modeled by active learning in our future work.

Acknowledgements

The research was funded by the K.U.Leuven research grant OT/09/028(VASI), the Natural Science Foundation of Jiangsu Province (BK20140071) and the Natural Science Foundation of China (61402519).

6. REFERENCES

[1] L. Boves, L. ten Bosch, and R. Moore, "Acorns - towards computational modeling of communication and recognition skills," in *International Conference on Cognitive Informatics*, 2007, pp. 349–356.

[2] J. Driesen and H. Van hamme, "Modelling vocabulary acquisition, adaptation and generalization in infants using adaptive bayesian plsa," *Neurocomputing*, vol. 74(11), pp. 1874–1882, 2011.

[3] I. Ayllon Clemente, M. Heckmann, and B. Wrede, "Incremental word learning: Efficient hmm initialization and large margin discriminative adaptation," *Speech Communication*, vol. 54, pp. 1029–1048, Nov. 2012.

[4] O. Räsänen and U. K. Laine, "A method for noise-robust context-aware pattern discovery and recognition from categorical sequences," *Pattern Recognition*, vol. 45, no. 1, pp. 606–616, 2012.

[5] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 186–197, 2008.

[6] A. Jansen, K. Church, and H. Hermansky, "Towards spoken term discovery at scale with zero resources," in *INTERSPEECH*, 2010, pp. 1676–1679.

[7] G.Aimetti, L.ten Bosch, and R.K.Moore, "The emergence of words: Modelling early language acquisition with a dynamic systems perspective," in *INTERSPEECH*, 2009.

[8] H Van hamme, "Hac-models: a novel approach to continuous speech recognition," in *INTERSPEECH*, 2008, pp. 2554–2557.

[9] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains," *Ann. Math. Statist.*, vol. 41(1), pp. 164–171, 1970.

[10] B. Vanluyten, J. Willems, and B. De Moor, "Structured nonnegative matrix factorization with applications to hidden markov realization and clustering," *Linear Algebra and Its Applications*, vol. 429, pp. 1409–1424, 2008.

[11] B. Lakshminarayanan and R. Raich, "Non-negative matrix factorization for parameter estimation in hidden markov models," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2010, pp. 89–94.

[12] G. Cybenko and V. Crespi, "Learning hidden markov models using non-negative matrix factorization," *IEEE Transactions on Information Theory*, vol. 57(6), pp. 3963–3970, 2011.

[13] M. Sun and H. Van hamme, "Joint training of non-negative tucker decomposition and discrete density hidden markov models," *Computer Speech and Language*, vol. 27(4), pp.969–988, 2013.

[14] B. H. Juang and L. R. Rabiner, "The segmental k-means algorithm for estimating parameters of hidden markov models," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 38, pp. 1639–1641, 1990.

[15] S. Ji, L. T. Watson, and L. Carin, "Semisupervised learning of hidden markov models via a homotopy method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 275–287, 2009.

[16] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77(2), pp. 257–286, 1989.

[17] C. Lew-Williams, B. Pelucchi, and J. R Saffran, "Isolated words enhance statistical language learning in infancy," *Developmental Science*, vol. 14(6), pp. 1323–1329, November 2011.

[18] V. Stouten, K. Demuynck, and H. Van hamme, "Discovering phone patterns in spoken utterances by non-negative matrix factorisation," *IEEE Signal Processing Letters*, vol. 15, pp. 131–134, 2008.